Coarse-graining of perplexity for the spatial distribution of molecules

Lili Xu [1], Chi Zhang,[2] Soho Oyama,[1] Manabu Machida [2,3,*] Tomoaki Kahyo,[1,4] and Mitsutoshi Setou [1,2,4]

[1]*Department of Cellular and Molecular Anatomy, Hamamatsu University School of Medicine, Hamamatsu 431-3192, Japan*
[2]*Department of Systems Molecular Anatomy, Institute for Medical Photonics Research, Preeminent Medical Photonics Education & Research Center, Hamamatsu University School of Medicine, Hamamatsu 431-3192, Japan*
[3]*Department of Informatics, Faculty of Engineering, Kindai University, Higashi-Hiroshima 739-2116, Japan*
[4]*International Mass Imaging Center, Hamamatsu University School of Medicine, Hamamatsu 431-3192, Japan*

Biological tissue consists of various molecules. Instead of focusing on a particular molecule, we consider the Shannon entropy which is calculated from the abundance of different molecules at each spot in the tissue. The spatial distribution of the Shannon entropy is of interest. In this paper, we first obtain the heat map of perplexity, whose logarithm is the entropy. To characterize the spatial variety of molecules, we propose a scalar k that is concerned with the coarse-graining of the perplexity heat map. To verify the usefulness of the number, experiments with mass spectrometry imaging were performed for mouse kidneys. We found that k has large values in the renal pelvis area, cortex area, veins, and arteries in the mouse kidney, whereas fractal dimensions fail to distinguish those regions.

I. INTRODUCTION

Among different molecules in biological tissue, quite often a target molecule is focused on and the position dependence of the abundance of the target molecule is investigated. An alternative approach to study the tissue with various molecules is to obtain a characteristic quantity such as the Shannon entropy and study its position dependence. In this paper, aimed at characterizing biological tissues from the viewpoint of the spatial distribution of the entropy, we perform coarse-graining.

Entropy has been used to analyze information of a medium [1]. For classical and quantum gases, fractal dimensions and their relation to entropy were investigated [2]. The Shannon entropy was used for avoided crossings in quantum chaos [3]. The concept of the Shannon entropy was used for the kinetics of colloidal particles [4]. See, also, [5,6]. Entropy was used for the study of quantum many-body systems [7]. An extension of the Shannon entropy was explored for biological diversity [8].

To study biological tissue from an informational point of view, heat maps of the Shannon entropy calculated from the abundance of molecules in each spot have been investigated [9,10]. In this paper, the characterization of the position dependence of the Shannon entropy is explored.

Mass spectrometry imaging (MSI) is a technique with ionizing molecules in a sample to provide both molecule species and positions of those molecules in the sample [11–13]. The resolution of an image is of the order of micrometer or sub-nanometer. At each spot of the sample, hundreds of m/z peaks are obtained.

In mass spectrometry, the Shannon entropy was calculated for m/z spectra [14–17]. See [18,19] for more details. Recently, the binning effect of the Shannon entropy [20], the relation between the Shannon entropy of mass spectra and molecules such as peptides and proteins [21], and a data-targeted extraction method for metabolite annotation [22] were investigated.

The Shannon entropy has been viewed as a physical quantity which gives information as to how molecules spatially vary. In [9], the Shannon entropy has been revisited to capture information from all peaks in mass spectra over the sample and a method based on the information entropy (Shannon entropy) for time-of-flight secondary ion mass spectrometry (TOF-SIMS) was proposed. Furthermore it was shown that without peak identification, the spatial distribution (heat maps) of the Shannon entropy of spectra indicates differences in materials and changes in the conditions of a material in a sample. The spatial distribution of the Shannon entropy was also studied for the matrix-assisted laser desorption-ionization (MALDI) MSI and a method to select candidate peaks was proposed [10].

In this study, we derive perplexity from entropy and develop an approach to visualize the spatial and mass-spectral diversities by coarse-graining. To this end, a slope k (see below) is introduced.

The remainder of the paper is organized as follows. In Sec. II, we introduce perplexity and define k. Section III gives the experimental results for MSI. Fractal dimensions are considered in Sec. IV. Section V is devoted to discussion. The conclusions are given in Sec. VI.

II. PERPLEXITY AND COARSE-GRAINING

We begin by introducing the Shannon entropy using intensities in the mass spectrum [9,10]. Let n be the number of

---

*machida@hiro.kindai.ac.jp
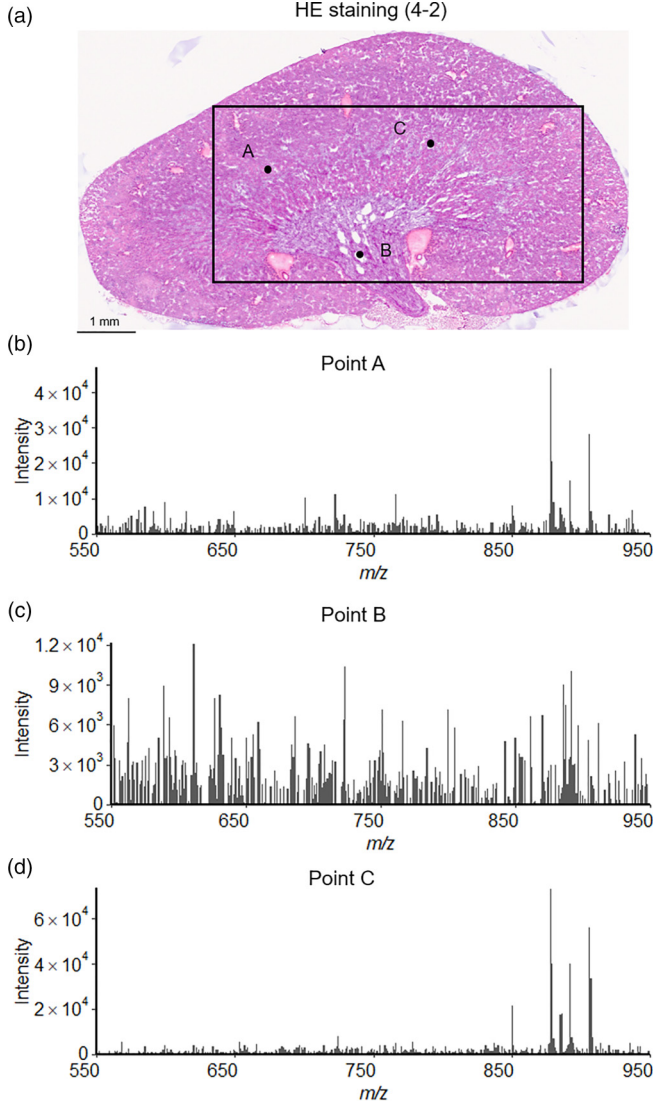
(a)

(b)

(c)

(d)

FIG. 1. A mouse kidney (sample 4-2) and examples of peaks on mass spectra at three different spots in the kidney.

intensities in the mass spectrum. In MSI, different intensities appear as a function of m/z at each pixel on the sample. Examples of such peaks are shown in Fig. 1 for the mouse kidney of sample 4-2 (see below). We call nonzero intensities the peaks. Each peak corresponds to an ion. Here, the dimensionless unit m/z is the ratio of the mass of an ion divided by the unified atomic mass unit (1 Dalton) and the absolute charge number (the absolute value of an integer number of elementary charges that an ion has gained or lost). Let p_i be the ith intensity in the mass spectrum which is normalized as $\sum_{i=1}^n p_i = 1$. In Fig. 1, the numbers of the peaks are 558, 478, 655 for points A,B,C, respectively.

We draw x and y axes on the image. At pixel (x, y), we define the Shannon entropy $H(x, y)$ as

$$H(x, y) = -\sum_{i=1}^n p_i(x, y) \log_2 p_i(x, y). \quad (1)$$

After plotting $H(x, y)$, a way to study the structure of $H(x, y)$ in the neighborhood of the point (x, y) is to consider coarse-graining. As described below, a computable scalar can be

obtained with the coarse-graining process for perplexity, which is introduced below, rather than the direct application of coarse-graining to the heat map $H(x, y)$.

Let us consider perplexity, which is an index to measure diversity. See [23] for an application of perplexity in biology. Perplexity (PP) is defined as [24]

$$\mathrm{PP}(x, y) = 2^{H(x,y)}. \quad (2)$$

Suppose that we randomly pick a peak in the mass spectrum with the probability given by the relative height of the peak. Then we consider the number of tries that is necessary for the particular peak to be selected. Indeed, the expected number of trials is the reciprocal of the probability for the peak. The perplexity is the weighted geometrical mean of the expected numbers of trials for all peaks in the mass spectrum. In this way, the perplexity provides a measure of diversity of peaks, which correspond to different molecules. Since $0 \leqslant H(x, y) \leqslant \log_2 n$, we have $1 \leqslant \mathrm{PP}(x, y) \leqslant n$.

Let us consider coarse-graining for MSI. We select a square region on the image and divide the square into $N(\varepsilon)$ subsquares with side length ε. Treating these $N(\varepsilon)$ subsquares as new pixels, relative peak intensities in the pixel at (x, y) can be expressed as $p_i^{(\varepsilon)}(x, y)$. We take averages in mass spectra for coarse-grained pixels. If the pixel size gets doubled by coarse-graining, we have $p_i^{(2\varepsilon)}(x, y) = \frac{1}{4}[p_i^{(\varepsilon)}(x_1, y_1) + p_i^{(\varepsilon)}(x_2, y_2) + p_i^{(\varepsilon)}(x_3, y_3) + p_i^{(\varepsilon)}(x_4, y_4)]$, where $(x_1, y_1) = (x, y)$, $(x_2, y_2) = (x + \varepsilon, y)$, $(x_3, y_3) = (x, y + \varepsilon)$, and $(x_4, y_4) = (x + \varepsilon, y + \varepsilon)$. We note that the pixels of size ε at (x_j, y_j) ($j = 1, 2, 3, 4$) are contained in the pixel of size 2ε at (x, y). Thus, the perplexity depends on ε and we can write $\mathrm{PP}^{(\varepsilon)}(x, y)$.

In addition to $\mathrm{PP}^{(\varepsilon)}(x, y)$ itself, we consider how $\mathrm{PP}^{(\varepsilon)}(x, y)$ behaves as ε varies by coarse-graining. We define $k(x, y)$ as

$$k(x, y) = \lim_{\varepsilon \to 0} \frac{\mathrm{PP}^{(\varepsilon)}(x, y)}{\ln \varepsilon}. \quad (3)$$

This $k(x, y)$ contains the information on how $H(x, y)$ behaves in the neighborhood of the point (x, y). Since we have found a linear dependence between $\ln \varepsilon$ and the perplexity (see below), the calculation of the limit in (3) is feasible for experimental data. Indeed, (3) means a power-law behavior of e^{PP} for small $\varepsilon > 0$ as

$$e^{\mathrm{PP}} \sim \varepsilon^k. \quad (4)$$

Below, we will see that the relation (4) holds for experimental data. In terms of the Shannon entropy, the relation (4) implies a power-law dependence of the double exponential of H on ε,

$$\exp(e^H) \sim \varepsilon^k. \quad (5)$$

Thus, k characterizes the local behavior of the spatial distribution of H.

III. EXPERIMENTS

A. Setup

We used C57BL/6 J female mice, with two four-month-old mice kidneys (samples 4-1 and 4-2) for MALDI MSI observation (using a high-resolution microscopic imaging mass spectrometer, iMScope). Sample preparation steps including mouse sacrifice, sample pre-reservation, and matrix spraying
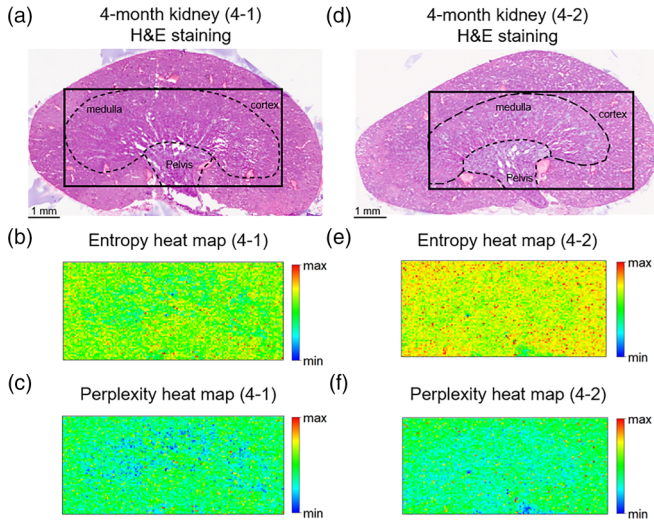
FIG. 2. Entropy and perplexity heat maps on kidney MSI data. (a),(d) HE staining of kidneys (samples 4-1 and 4-2) of four-month-old mice and ROIs (marked by rectangles). (b),(e) Entropy heat maps in ROIs. In the ROI of sample 4-1, the entropy changed between 5.0 and 9.6. In the ROI of sample 4-2, the entropy changed between 7.2 and 9.7. (c),(f) Perplexity heat maps in ROIs. In the ROI of sample 4-1, the perplexity changed between 149.0 and 846.5. In the ROI of sample 4-2, the perplexity changed between 32.2 and 755.3. Scale bar: 1 mm.

followed our previous paper [10]. The experimental conditions were as follows: negative ion mode, m/z range between 550 and 950, laser strength of 45%, and number of irradiations of 100. All experiments in this study were performed in compliance with the licensing instructions from the Institutional Animal Care and Use Committees of Hamamatsu University, School of Medicine, Japan (permission code 2015028).

For sample 4-1, $n = 3246$. For sample 4-2, $n = 3584$.

B. Perplexity heat maps

First, we select a region of interest (ROI). Figures 2(a) and 2(d) show the anatomical structure through hematoxylin-eosin (HE) staining. We calculated the Shannon entropy and perplexity on each spot of the kidney MSI data and plotted heat maps. In Figs. 2(b) and 2(c) (sample 4-1) and Figs. 2(e) and 2(f) (sample 4-2), there are differences between entropy and perplexity heat maps. Moreover, sample dependencies can be seen in the comparisons between Figs. 2(b) and 2(e) and between Figs. 2(c) and 2(f). In Fig. 2(c), blue and green spots (i.e., low and middle perplexity spots) dominate in the medulla and cortex areas, respectively. In Fig. 2(e), the pelvis and medulla areas (green) (i.e., middle perplexity areas) can be distinguished from the cortex (yellow) (i.e., high perplexity spots). For both of the samples 4-1 and 4-2, the medulla area has lower entropy values than the cortex area.

C. The slope k

To investigate the spatial distribution of perplexity, we picked several points on the image for the kidney of sample 4-2 [Figs. 3(a) and 3(b)]. We found a linear trend in the plot of $\ln \varepsilon$ and perplexity for small $\ln \varepsilon$. In Fig. 2,
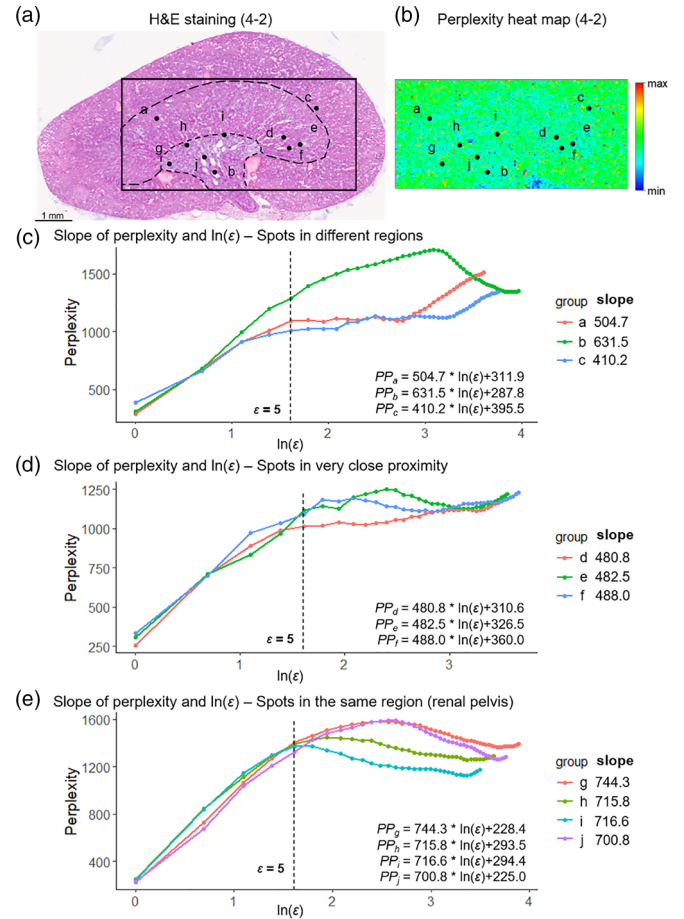

FIG. 3. (a),(b) The HE staining and perplexity heat map of the four-month-old mouse kidney (sample 4-2) and spots selected for demonstrating slope calculations. The range of the perplexity was between 32.2 and 755.3. (c) Semilogarithmic plots of $\ln \varepsilon$ and the perplexity. For spots a–c, slopes were 504.7, 631.5, and 410.2, respectively. (d) For spots d–f, slopes were 480.8, 482.5, and 488.0, respectively. (e) For spots g–j, slopes were 744.3, 715.8, 716.6, and 700.8, respectively. Scale bar: 1 mm.

the unit of ε is the length of the original pixels. Different spots (a–c) were chosen and the linear dependence was found at each spot with different slopes (a = 504.7, b = 631.5, and c = 410.2) [Fig. 3(c)]. Spots in the proximity (d–f) showed similar slopes (d = 480.8, e = 482.5, and f = 488.0) [Fig. 3(d)]. Spots in the same region (g–j) presented similar slopes (g = 744.3, h = 715.8, i = 716.6, and j = 700.8) [Fig. 3(e)]. Thus, k can be experimentally determined.

Figures 4(a) and 4(b) show heat maps of $k(x, y)$. We found that the renal pelvis area has large k (red and yellow spots), while k is small (blue and green spots) in the medulla area [Figs. 4(c) and 4(d)]. There were green and yellow spots in the cortex area. We also found pixels of large k in the areas denoted by * and ** in Figs. 4(c) and 4(d). As examples, we compare the HE images and heat maps in these rectangular areas. The magnified figures are shown in Figs. 4(e) and 4(f). We discovered that large k appears at the vein and artery.
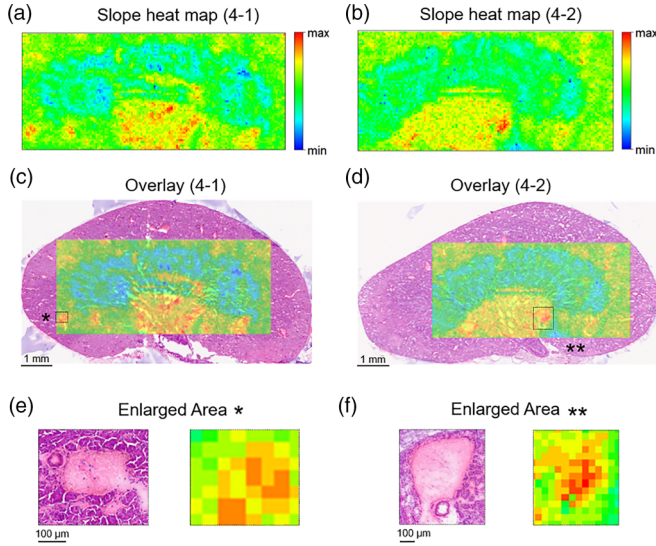
FIG. 4. Heat maps of the slope k are shown with overlays with HE staining. (a),(b) Heat maps of k for samples 4-1 and 4-2. In (a) and (b), [min, max] is [171.1,1105.4] and [104.4,902.6], respectively. (c),(d) Overlays of the heat map and HE staining of the kidneys. Scale bar: 1 mm. (e),(f) Enlarged areas for the HE staining and heat map of k. Scale bar: 10 μm.

IV. FRACTAL DIMENSIONS

Here we take the unit of ε to be the side length of the original pixels. Let N_0 be the number of original pixels in the ROI. The number of pixels after binning is $N(\varepsilon) = N_0/\varepsilon^2$. Fractal dimension D_0 is calculated as $D_0 = -\lim_{\varepsilon \to 0} \ln N(\varepsilon)/\ln \varepsilon$ [25]. From the definition, we see that in this case, $D_0 = 2$ exactly. For given ε, let us introduce

$$P_i(\varepsilon) = \frac{\sum_{(x,y) \in i\text{th pixel}} H(x, y)}{\sum_{(x,y) \in \text{ROI}} H(x, y)}, \qquad (6)$$

where $H(x, y)$ is the Shannon entropy which is computed using original pixels. The number of pixels depends on ε; the number decreases as ε becomes large.

Similar to D_0, the information dimension D_1 and correlation dimension D_2 [26,27] are given by $D_1 = \lim_{\varepsilon \to 0} \sum_{i=1}^{N(\varepsilon)} P_i(\varepsilon) \ln P_i(\varepsilon)/\ln \varepsilon$, $D_2 = \lim_{\varepsilon \to 0} \ln[\sum_{i=1}^{N(\varepsilon)} P_i(\varepsilon)^2]/\ln \varepsilon$. In general, we have $D_q = \frac{1}{q-1} \lim_{\varepsilon \to 0} [\ln Z_q(\varepsilon)]/(\ln \varepsilon)$, where $Z_q(\varepsilon) = \sum_{i=1}^{N(\varepsilon)} P_i(\varepsilon)^q$.

In addition to the fractal dimension D_0, the information dimension D_1 and the correlation dimension D_2 were calculated at different points in the images. We found $D_1 = 2.0$ and $D_2 = 2.0$ for both kidneys. This means that the fractal nature is not found for $q = 0, 1, 2$.

V. DISCUSSION

We found that the relation $e^{\text{PP}} \sim \varepsilon^k$ in (4) holds for the experimental data. Although this relation stems from the statistical nature of the molecules, it is still an open problem of how k reflects the spatial distribution of molecules in the sample. The physical and biological reasons of the power-law behavior need to be clarified in the future.

We note that $k(x, y)$ is positive even when the Shannon entropy for pixels in the neighborhood of (x, y) is unchanged if spectral patterns in the neighborhood have a variety. That is, k has different values even if the heat map is homogeneous. In this sense, k is more informative than entropy and perplexity.

Different anatomical regions showed different entropy and perplexity (Fig. 2). In normal kidneys, the cortex contains more histological structures including proximal tubules, glomeruli, cortical distal tubules, and interstitial structures, while the medulla contains interstitial structures, medullary tubules, and corticomedullary tubules [28]. Our detected m/z range (550–950) was mainly lipids. It is known that the lipid expression of histological structures within the cortex is similar, unlike the lipid expression within the medulla [28], which may explain the inconsistency of the entropy and perplexity in Fig. 1.

The fact that the entropy and perplexity in the medulla area were both lower than those in the cortex area in Fig. 2 implies less complexity of biomolecular and chemical information in the medulla area. The major lipids in the cortex are phospholipids, whereas the medulla is dominated by neutral lipids [29]. Phospholipids are composed of a head group, a glycerol backbone, and fatty acid chains [30]. Neutral lipids, on the other hand, contain only a glycerol backbone and fatty acid chains [31]. This might explain low-entropy spots in the medulla area.

When the mass spectra drastically vary in space (e.g., when different high-intensity peaks appear in neighboring spots), entropy grows by coarse-graining. In this case, perplexity rapidly increases and k at this spot becomes large. By contrast, if the mass spectra are spatially more or less similar, entropy changes mildly by coarse-graining and k is small.

In Fig. 4, k is large (red and yellow spots) around the vein and artery area in the kidney. The HE staining shows blood vessels. Indeed, pixels inside and outside the blood vessels have quite different mass spectra. On the other hand, k is small (blue spots) in the medulla area. This implies that the region of medulla is relatively homogeneous.

VI. CONCLUSIONS

We have proposed the use of perplexity and introduced k in coarse-graining. We found that the heat map of k reveals new structures which are not clearly visible in the Shannon-entropy heat map. Although the use of peaks in the mass spectrum as a distribution is not yet established, the spatial distribution of k will help characterize biological tissues.

Since experimental data show the power-law behavior of the exponential of perplexity in (4), it was possible to produce heat maps of k for kidneys. These heat maps elucidate structures of the Shannon entropy, while the information was not clearly extracted with fractal dimensions.

[1] R. Frigg and C. Werndl, Entropy – A guide for the perplexed, in *Probabilities in Physics*, edited by C. Beisbart and S. Hartmann (Oxford University Press, Oxford, 2011), pp. 115–142.

[2] F. Büyükkiliç and D. Demirhan, A fractal approach to entropy and distribution functions, Phys. Lett. A **181**, 24 (1993).

[3] K.-W. Park, S. Moon, Y. Shin, J. Kim, K. Jeong, and K. An, Shannon entropy and avoided crossings in closed and open quantum billiards, Phys. Rev. E **97**, 062205 (2018).

[4] D. Wu, K. Ghosh, M. Inamdar, H. J. Lee, S. Fraser, K. Dill, and R. Phillips, Trajectory approach to two-state kinetics of single particles on sculpted energy landscapes, Phys. Rev. Lett. **103**, 050603 (2009).

[5] E. T. Jaynes, The minimum entropy production principle, Annu. Rev. Phys. Chem. **31**, 579 (1980).

[6] W. T. Grandy, Principle of maximum entropy and irreversible processes, Phys. Rep. **62**, 175 (1980).

[7] B. Tarighi, R. Khasseh, M. N. Najafi, and M. A. Rajabpour, Universal logarithmic correction to Rényi (Shannon) entropy in generic systems of critical quadratic fermions, Phys. Rev. B **105**, 245109 (2022).

[8] C. Ricotta, Parametric scaling from species relative abundances to absolute abundances in the computation of biological diversity: A first proposal using Shannon's entropy, Acta Biotheor. **51**, 181 (2003).

[9] S. Aoyagi, K. Mizomichi, K. Kamochi, and A. Miisho, Interpretation of TOF-SIMS data based on information entropy of spectra, Surf. Interface Anal. **54**, 356 (2022).

[10] L. Xu, K. Kikushima, S. Sato, A. Islam, T. Sato, S. Aramaki, C. Zhang, T. Sakamoto, F. Eto, Y. Takahashi, I. Yao, M. Machida, T. Kahyo, and M. Setou, Spatial distribution of the Shannon entropy for mass spectrometry imaging, PLoS ONE **18**, e0283966 (2023).

[11] S. Shimma, Y. Sugiura, T. Hayasaka, Y. Hoshikawa, T. Noda, and M. Setou, MALDI-based imaging mass spectrometry revealed abnormal distribution of phospholipids in colon cancer liver metastasis, J. Chromatogr. B **855**, 98 (2007).

[12] K. Chughtai and R. M. A. Heeren, Mass spectrometric imaging for biomedical tissue analysis, Chem. Rev. **110**, 3237 (2010).

[13] B. Spengler, Mass spectrometry imaging of biomolecular information, Anal. Chem. **87**, 64 (2015).

[14] A. G. Ferrige, M. J. Seddon, S. Jarvis, J. Skilling, and R. Aplin, Maximum entropy deconvolution in electrospray mass spectrometry, Rapid Commun. Mass Spectrom. **5**, 374 (1991).

[15] B. B. Reinhold and V. N. Reinhold, Electrospray ionization mass spectrometry: Deconvolution by an entropy-based algorithm, J. Am. Soc. Mass Spectrom. **3**, 207 (1992).

[16] A. Broersen, R. van Liere, A. F. M. Altelaar, R. M. A. Heeren, and L. A. McDonnell, Automated, feature-based image alignment for high-resolution imaging mass spectrometry of large biological samples, J. Am. Soc. Mass Spectrom. **19**, 823 (2008).

[17] W. M. Abdelmoula, K. Škrášková, B. Balluff, R. J. Carreira, E. A. Tolner, B. P. F. Lelieveldt, L. van der Maaten, H. Morreau, A. M. J. M. van den Maagdenberg, R. M. A. Heeren, L. A. McDonnell, and J. Dijkstra, Automatic generic registration of mass spectrometry imaging data to histology using nonlinear stochastic embedding, Anal. Chem. **86**, 9204 (2014).

[18] I. A. Kaltashov and S. J. Eyles, *Mass Spectrometry in Biophysics: Conformation and Dynamics of Biomolecules* (Wiley, New York, 2005).

[19] S. Aoyagi, Review of TOF-SIMS bioanalysis using mutual information, Surf. Interface Anal. **41**, 136 (2009).

[20] R. M. T. Madiona, D. L. J. Alexander, D. A. Winkler, B. W. Muir, and P. J. Pigram, Information content of ToF-SIMS data: Effect of spectral binning, Appl. Surf. Sci. **493**, 1067 (2019).

[21] Y. Li, T. Kind, J. Folz, A. Vaniya, S. S. Mehta, and O. Fiehn, Spectral entropy outperforms MS/MS dot product similarity for small-molecule compound identification, Nat. Methods **18**, 1524 (2021).

[22] F. Zheng, L. You, W. Qin, R. Ouyang, W. Lv, L. Guo, X. Lu, E. Li, X. Zhao, and G. Xu, MetEx: A targeted extraction strategy for improving the coverage and accuracy of metabolite annotation in liquid chromatography-high-resolution mass spectrometry data, Anal. Chem. **94**, 8561 (2022).

[23] L. Jost, Entropy and diversity, Oikos **113**, 363 (2006).

[24] K. P. Nelson, Reduced perplexity: Uncertainty measures without entropy, in *Recent Innovations in Info-metrics: A Cross-disciplinary Perspective on Information and Information Processing* (Oxford University Press, Oxford, 2014).

[25] S. Kak, Information theory and dimensionality of space, Sci. Rep. **10**, 20733 (2020).

[26] A. Rényi, On the dimension and entropy of probability distributions, Acta Math. Acad. Sci. Hung. **10**, 193 (1959).

[27] K. J. Falconer, *Fractal Geometry: Mathematical Foundations and Applications* (Wiley, Chichester, England, 2003).

[28] L. Martín-Saiz, L. Mosteiro, J. D. Solano-Iturri, Y. Rueda, J. Martín-Allende, I. Imaz, I. Olano, B. Ochoa, O. Fresnedo, J. A. Fernández, and G. Larringa, High-resolution human kidney molecular histology by imaging mass spectrometry of lipids, Anal. Chem. **93**, 9364 (2021).

[29] R. E. Druilhet, M. L. Overturf, and W. M. Kirkendall, Cortical and medullary lipids of normal and nephrosclerotic human kidney, Intl. J. Biochem. **9**, 729 (1978).

[30] S. Drescher and P. van Hoogevest, The phospholipid research center: Current research in phospholipids and their use in drug delivery, Pharmaceutics **12**, 1235 (2020).

[31] K. Athenstaedt and G. Daum, The life cycle of neutral lipids: Synthesis, storage and degradation, Cell. Mol. Life Sci. **63**, 1355 (2006).